

Contents

Preface	xi
1 Introduction	1
1.1 Book overview	2
1.2 Overview of <i>omic</i> data	2
1.2.1 Genomic data	3
1.2.1.1 Genomic SNP data	3
1.2.1.2 SNP arrays	3
1.2.1.3 Sequencing methods	4
1.2.2 Genomic data for other structural variants	5
1.2.3 Transcriptomic data	5
1.2.3.1 Microarrays	6
1.2.3.2 RNA-seq	6
1.2.4 Epigenomic data	7
1.2.5 Exposomic data	8
1.3 Association studies	8
1.3.1 Genome-wide association studies	9
1.3.2 Whole transcriptome profiling	10
1.3.3 Epigenome-wide association studies	10
1.3.4 Exposome-wide association studies	11
1.4 Publicly available resources	12
1.4.1 dbGaP	12
1.4.2 EGA	13
1.4.3 GEO	13
1.4.4 1000 Genomes	14
1.4.5 GTEx	14
1.4.6 TCGA	15
1.4.7 Others	15
1.5 Bioconductor	16
1.5.1 R	16
1.5.2 <i>Omic</i> data in Bioconductor	17
1.6 Book's outline	18

2	Case examples	21
2.1	Chapter overview	21
2.2	Reproducibility: The case for public data repositories	21
2.3	Case 1: dbGaP	22
2.4	Case 2: GEO	30
2.5	Case 3: GTEx	37
2.6	Case 4: TCGA	40
2.7	Case 5: NHANES	45
3	Dealing with omic data in Bioconductor	49
3.1	Chapter overview	49
3.2	<i>snpMatrix</i>	50
3.3	<i>ExpressionSet</i>	52
3.4	<i>SummarizedExperiment</i>	53
3.5	<i>GRanges</i>	55
3.6	<i>RangedSummarizedExperiment</i>	62
3.7	<i>ExposomeSet</i>	64
3.8	<i>MultiAssayExperiment</i>	67
3.9	<i>MultiDataSet</i>	69
4	Genetic association studies	75
4.1	Chapter overview	75
4.2	Genetic association studies	76
4.2.1	Analysis packages	76
4.2.2	Association tests	77
4.2.3	Single SNP analysis	79
4.2.4	Hardy–Weinberg equilibrium	83
4.2.5	SNP association analysis	84
4.2.6	Gene $\times$ environment and gene $\times$ gene interactions	91
4.3	Haplotype analysis	94
4.3.1	Linkage disequilibrium heatmap plots	95
4.3.2	Haplotype estimation	98
4.3.3	Haplotype association	98
4.3.4	Sliding window approach	99
4.4	Genetic score	102
4.5	Genome-wide association studies	107
4.5.1	Quality control of SNPs	109
4.5.2	Quality control of individuals	110
4.5.3	Population ancestry	116
4.5.4	Genome-wide association analysis	117
4.5.5	Adjusting for population stratification	119
4.6	Post-GWAS visualization and interpretation	126
4.6.1	Genome-wide associations for imputed data	129

5 Genomic variant studies	133
5.1 Chapter overview	133
5.2 Copy number variants	134
5.2.1 CNV calling	136
5.3 Single CNV association	153
5.3.1 Inferring copy number status from signal data	157
5.3.2 Measuring uncertainty of CNV calling	161
5.3.3 Assessing the association between CNVs and traits	161
5.3.3.1 Modeling association	162
5.3.3.2 Global test of associations	164
5.3.4 Whole genome CNV analysis	166
5.4 Genetic mosaicisms	169
5.4.1 Calling genetic mosaicisms	169
5.4.2 Calling the loss of chromosome Y	178
5.5 Polymorphic inversions	188
5.5.1 Inversion detection	189
5.5.2 Inversion calling	191
5.5.3 Inversion association	195
6 Addressing batch effects	199
6.1 Chapter overview	199
6.2 SVA	200
6.3 ComBat	206
7 Transcriptomic studies	211
7.1 Chapter overview	211
7.2 Microarray data	212
7.2.1 Normalization	212
7.2.2 Filter	216
7.2.3 Differential expression	220
7.3 Next generation sequencing data	225
7.3.1 Normalization	229
7.3.2 Gene filtering	235
7.3.3 Differential expression	235
8 Epigenomic studies	245
8.1 Chapter overview	245
8.2 Epigenome-wide association studies	245
8.3 Methylation arrays	246
8.4 Differential methylation analysis	247
8.5 Methylation analysis of a target region	254
8.6 Epigenomic and transcriptomic visualization results	257
8.7 Cell proportion estimation	260

9 Exposomic studies	263
9.1 Chapter overview	263
9.2 The exposome	264
9.2.1 Exposomic data	265
9.3 Exposome characterization	266
9.4 Exposome-wide association analyses	275
9.5 Association between exposomic and other <i>omic</i> data	278
9.5.1 Exposome-transcriptome data analysis	279
9.5.2 Exposome-methylome data analysis	287
10 Enrichment analysis	291
10.1 Chapter overview	291
10.2 Enrichment analysis and statistical power	292
10.3 Gene set annotations	293
10.4 Over representation analysis	296
10.5 Overlap with functional genomic regions	307
10.6 Chemical and environmental enrichment	310
11 Multiomic data analysis	315
11.1 Chapter overview	315
11.2 Multiomic data	316
11.3 Massive pair-wise analyses between <i>omic</i> datasets	316
11.4 Multiple- <i>omic</i> data integration	322
11.4.1 Multi-staged analysis	323
11.4.1.1 Genomic variation analysis	323
11.4.2 Domain-knowledge approach	328
11.4.3 Meta-dimensional analysis	332
11.4.3.1 Principal component analysis	332
11.4.3.2 Sparse principal component analysis	336
11.4.3.3 Canonical correlation and coinertia analyses	339
11.4.3.4 Regularized generalized canonical correlation	345
Bibliography	353
Index	371